Создание и сбор полнотекстовых электронных ресурсов в университетской библиотеке

Е. А. Негуляев, Е. А. Охезина
Уральский государственный университет

Сегодня можно с уверенностью сказать, что использование электронных ресурсов из «модной» проблемы превратилось в повседневную практику библиотек. При этом перед библиотеками встают задачи, которые мало чем отличаются от традиционных: определение источников комплектования, каталогизация, организация доступа. Только теперь все они решаются для нового вида информационных ресурсов.

Существует несколько источников получения электронных ресурсов для библиотек. Это подписка на коммерческие полнотекстовые базы данных, покупка электронных ресурсов на твердых копиях (CD-ROM и т. п.), самостоятельное создание или сбор электронных ресурсов и интеграция в библиотечное обслуживание веб-ресурсов.

Первые два подхода стали уже привычными, трудно найти крупную российскую библиотеку, которая не была бы подписана на коммерческие базы данных, или по крайней мере не приняла участие в бесплатном тестовом доступе [1] . Это же относится и к покупке энциклопедий, полнотекстовых изданий, архивов журнальных статей и т. п. на CD-ROM.

Остальные способы, если и не являются экзотикой, то встречаются не так уж часто, во многих библиотеках они до сих пор не вышли из экспериментальной стадии. Любой обмен опытом в этой области чрезвычайно полезен.

Электронные коллекции дореволюционных изданий

Научная библиотека Уральского государственного университета (НБ УрГУ) приступила к самостоятельному созданию полнотекстовых электронных ресурсов в январе 2001 года. К этому времени в библиотеке была накоплена первичная техническая база, позволившая организовать процессы оцифровки изданий и выдачи электронных копий читателям.

Изначально во главу угла был поставлен принцип эффективности создаваемых электронных коллекций — оцифровка изданий должна была проводиться не для накопления абстрактных показателей в виде количества страниц и изданий и не для выполнения грантовых обязательств, а для решения насущных проблем библиотеки.

Первой такой проблемой стала повышенная эксплуатация некоторых видов изданий. В связи с изменением стандартов образования резко увеличилась выдача книг, изданных во второй половине XIX — начале XX вв. (издания в дореформенной орфографии), что заставило вплотную задуматься над проблемами их физической сохранности. Оцифровка изданий представлялась как один из возможных способов одновременного решения проблем расширения доступа и обеспечения физической сохранности в том случае, если большинству читателей будет предоставляться цифровая копия издания, а не физический экземпляр [2].

Технологический процесс оцифровки должен был учитывать реалии нашей библиотеки: от имеющегося оборудования до квалификации персонала, — при этом обеспечивая высокую производительность труда. Первой нашей находкой стало использование графического формата DjVu — довольно экзотичного для 2001 года решения. DjVu позволял создавать точные графические копии книжных страниц с высоким разрешением (стандартно в 300 dpi), сжимая их до 20–30 Kb. Единственным недостатком была невозможность сохранять в DjVu файле распознанный текст на русском языке, но этот недостаток не был для нас актуальным: качество распознавания текстов в дореформенной орфографии недостаточно для использования без их дополнительной обработки.

Впрочем, мы решили и некоторые проблемы с распознаванием текстов в дореформенной орфографии. Нам удалось построить технологический процесс, который обеспечивает 90% верно воспроизведенных слов при одновременном переводе их из дореформенной в современную орфографию [3]. Заметим, что этот результат достигается полностью автоматизированными методами. При необходимости оставшиеся ошибки могут быть исправлены вручную, таким образом у нас обработано несколько дореволюционных изданий. Работа над распознаванием дореволюционных изданий и переводом распознанных текстов в современную орфографию преследует одну цель — сделать этот материал пригодным для полнотекстового поиска. Для представления в чисто текстовом виде 10% или даже 1% ошибок — заметная величина [4].

Первыми объектами сканирования стали самые востребованные дореволюционные издания из фондов Отдела редких книг нашей библиотеки. Среди них оказались классические издания по правоведению, истории русского права, исторические исследования и фундаментальные публикации исторических источников. Это дало возможность сформировать нам две тематические коллекции: коллекцию сочинений по истории и коллекцию правоведческих книг, названную нами «Правовая история России». Дальнейшее развитие оцифровки дореволюционных изданий будет идти с учетом пополнения этих тематических коллекций.

Постоянно работой по оцифровке изданий был занят один человек, работавший в библиотеке на 0,5 ставки. До октября 2002 года нам удалось довести объем цифровой коллекции дореволюционных книг до 25 тыс. страниц. Главный критерий отбора — востребованность изданий читателями — себя полностью оправдал: оказалось, что цифровая коллекция редких изданий покрывает около 10% общего количества книговыдач Отдела редких книг. В результате мы получили не только ожидаемую выгоду — сохранение физического состояния бумажных экземпляров, сверх нее мы сократили выдачу и расстановку книг, которые в любой библиотеке являются одними из наиболее трудоемких операций.

Оказалось, что наличие электронных копий редких изданий выгодно и большинству наших читателей. Очень востребованной оказалась услуга записи фрагментов оцифрованных изданий на CD-R, позволяющая читателям легко получить копии необходимых страниц.

Коллекция авторефератов и диссертаций

Следующим направлением работы, за реализацию которого мы взялись летом 2001 года, стало создание цифровой коллекции авторефератов диссертаций. Здесь на первое место вышли не технологические вопросы, а проблемы авторского права на современные издания и организационные проблемы. Важно было отработать технологическую цепочку: Ученый совет—Автор—Библиотека—Цифровая копия—Доступ к цифровой копии. Основу нашей коллекции составляют авторефераты и диссертации, защищенные в Ученом совете университета. Диссертанту после успешной защиты предлагается подписать договор на передачу неисключительных прав на публикацию электронной версии автореферата и/или диссертации. Автор добровольно определяет, что будет являться предметом договора (автореферат и диссертация, автореферат или диссертация) и устанавливает уровень доступа к цифровой копии. Наша задача заключается в том, чтобы объяснить диссертанту насколько полезной будет для него публикация электронной версии его научной работы. Опыт показал, что публикация электронной версии автореферата всегда приветствуется авторами, особенно после демонстрации возможностей доступа к цифровой коллекции через сайт библиотеки и других поисковых систем. Что касается публикации цифровой копии полного текста диссертации, то тут мы предлагаем автору подумать, открыто говоря о всех «плюсах» и «минусах» этого решения. Так как практически все авторы располагают электронным вариантом своей работы, то от библиотеки требуется лишь преобразовать его к общему формату хранения (в нашем случае это PDF), каталогизировать его, прописать в файле необходимые метаданные и обеспечить доступ к файлу.

В настоящее время коллекция насчитывает 186 авторефератов и 10 диссертаций, общим объемом около 7 тыс. страниц. Если бы библиотека не проявила инициативу и не организовала сбор этих информационных ресурсов, большинство электронных версий были бы просто утрачены для пользователей.

В настоящее время мы думаем над упрощением цепочки взаимоотношений Ученый совет—Автор—Библиотека таким образом, чтобы автор диссертации получил возможность подписать договор и передать электронную копию своей работы библиотеке университета еще в Ученом совете. Следующим шагом может также стать пополнение коллекции за счет работ сотрудников университета, защитивших свои диссертации в других Ученых советах.

Описание полнотекстовых ресурсов

Нам пришлось существенно расширить библиографические описания авторефератов и диссертаций, создаваемые в формате MARC21. За основу был взят шаблон, разработанный Российской Государственной библиотекой в рамках проекта создания национальной электронной библиотеки авторефератов и диссертаций. Теперь в библиографических записях содержится информация о научном руководителе, оппонентах, дате защиты и другие сведения, увеличивающая их поисковые возможности. Следующая таблица дает представление о полях, добавленых для описания авторефератов и диссертаций:

№ поля MARC21	Значение поля
024	70 $3 oksvnk $a 10.02.01
502	00 $a Защищена 02.10.18
700	10 $4 sad $a Бабенко Л. Г. $c Д-р филол. наук $e Науч. конс.
700	10 $4 opn $a Кусова М. Л. $c Д-р филол. наук $e Оппонент
700	10 $4 opn $a Сивкова Т. Н. $c Канд. филол. наук $e Оппонент
710	20 $a Урал. гос. ун-т им. А. М. Горького $b Каф. соврем. рус. яз.
710	20 $a Урал. гос. пед. ун-т
710	20 $a Тамб. гос. ун-т

Информация, заносимая в поля 024 (номер специальности и код классификатора, заполняется по Общероссийскому классификатору специальностей высшей научной квалификации), 502 (дата защиты), 700 (Научный руководитель/консультант, оппонент, соавторы) и 710 (организации, в которых готовилась и защищалась диссертация), позволяет организовать дополнительные точки доступа для поиска.

Библиографическое описание в MARC-формате является основным источником сведений об документе. Кроме него в настоящий момент в каждом PDF файле с электронной копией документа обязательно заполняется 4 поля с метаинформацией: «Автор» (Author), «Заглавие» (Title), «Название

организации, опубликовавшей ресурс» (DC.Publisher.CorporateName), «Адрес организации, опубликовавшей ресурс» (DC.Publisher.CorporateName.Address). Первые два являются структурной частью любого PDF файла. Для двух последних используется возможность Acrobat создавать пользовательские наборы метаданных, мы заполняем их в соответствии с форматом Dublin Core. Мы рассматриваем включение метаинформации непосредственно в PDF файлы с электронными копиями исключительно с практической стороны: они позволяют получить минимальную информацию о документе в отрыве от электронного каталога и могут быть использованы в качестве уточняющих критериев при полнотекстовом поиске по коллекции документов (см. ниже). Необходимость расширения набора метаданных, включаемых непосредственно в сам PDF файл, в настоящий момент изучается.

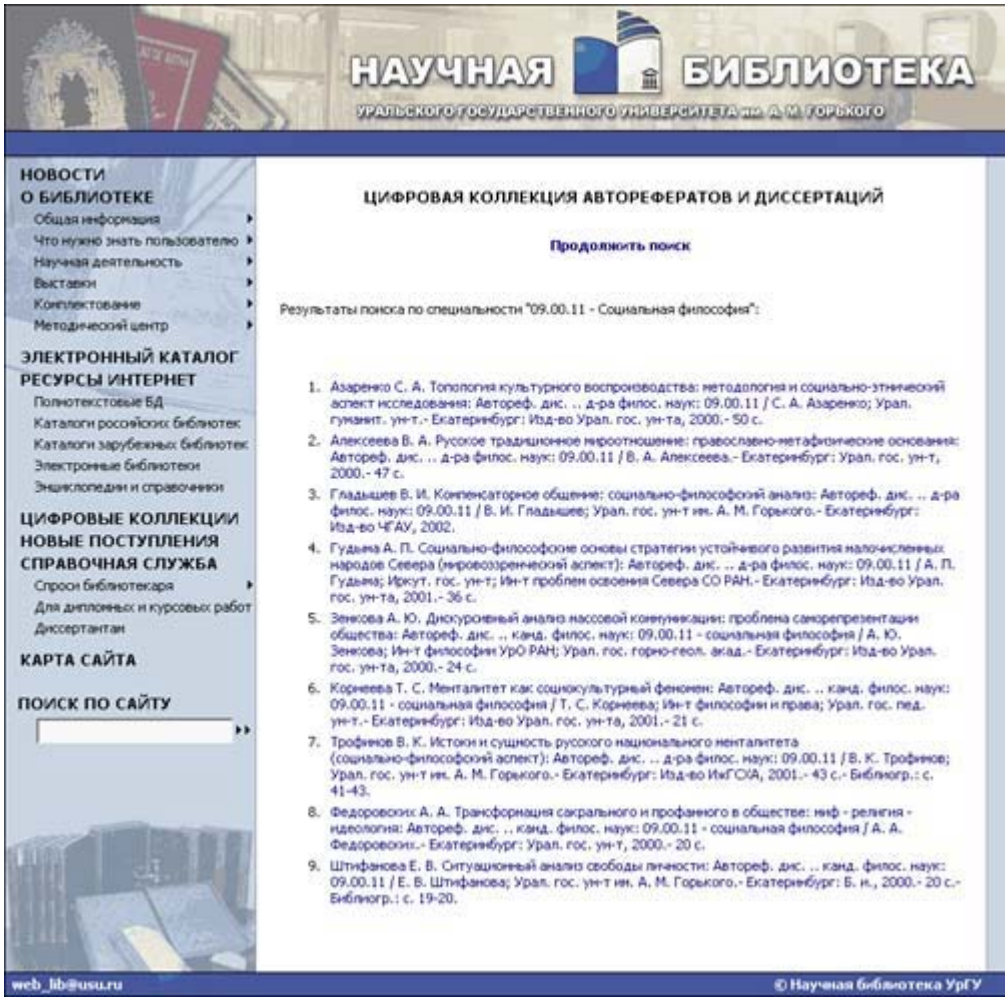
Доступ к цифровым коллекциям

Коллекция авторефератов и диссертаций в августе 2003 года была первой интегрирована в наш библиотечный сайт (http://lib.usu.ru/) [5]. В целом мы видим два основных способа доступа к цифровым копиям: доступ через сайт, где ресурсы располагаются по коллекционному и тематическому принципу, и доступ через электронный каталог библиотеки.

Каждый из этих способов обладает собственными достоинствами и недостатками, а наличие двух интерфейсов позволяет пользователю выбрать наиболее удобный.

В случае доступа через электронный каталог цифровая копия становится доступной по URL, записываемому в стандартное для всех форматов семейства MARC поле 856. Каталог НБ УрГУ доступен по протоколу Z39.50, что позволяет интегрировать его с другими аналогичными каталогами, расширяя тем самым и возможности доступа к полнотекстовой информации.

Специализированный тематический доступ к цифровым коллекциям на библиотечном сайте является упрощенным по сравнению с доступом через электронный каталог. Но при этом — более быстрым и удобным для пользователя. Использование этого вида доступа помогает решить еще одну задачу — сделать доступными полные тексты коллекций для поисковых систем, что практически невозможно при доступе только через электронный каталог [6].



Доступ к коллекции авторефератов и диссертаций на сайте библиотеки УрГУ

Создание цифровых коллекций учебных материалов

Накопленный за 2,5 года опыт работы над созданием электронных копий и организацией доступа к ним позволил нам летом 2003 года поставить задачу формирования больших коллекций учебных материалов.

Суть проекта формулировалась следующим образом: студент УрГУ должен иметь возможность получить на руки компакт-диск, на котором будут представлены все материалы, необходимые для изучения курса: методические пособия, учебники, научные монографии, хрестоматии, статьи из периодических изданий. Объем материалов для одного курса оценивался приблизительно в 15 тыс. страниц.

При подготовке проекта была поставлена задача обеспечить полнотекстовый поиск по созданным учебным коллекциям, причем добавить возможность поиска с учетом информации в полях с метаданными, такими как имя автора, название произведения и т. п.

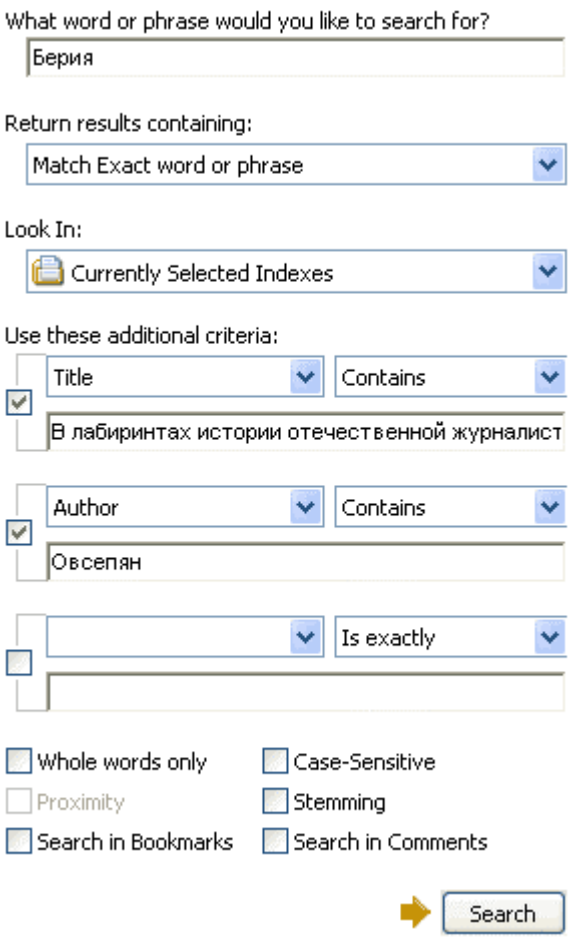
В качестве основы для представления цифровых копий был принят формат pdf. Причин для этого выбора было три: во-первых, pdf позволяет сохранять файл после распознавания в режиме «текст под изображением», а значит полностью исключить процедуру ручного исправления ошибок распознавания; во-вторых, средствами pdf достаточно легко можно организовать полнотекстовый поиск; в-третьих, возможности сжатия файлов в pdf достаточны для размещения на одном CD-R необходимого нам количества отсканированных страниц.

Следует отметить, что два последних пункта стали реальностью только с выходом шестой версии пакета Adobe Acrobat, увидевшей свет в конце мая 2003 г., поэтому для выполнения нашего проекта пришлось применить самые последние решения на основе pdf.

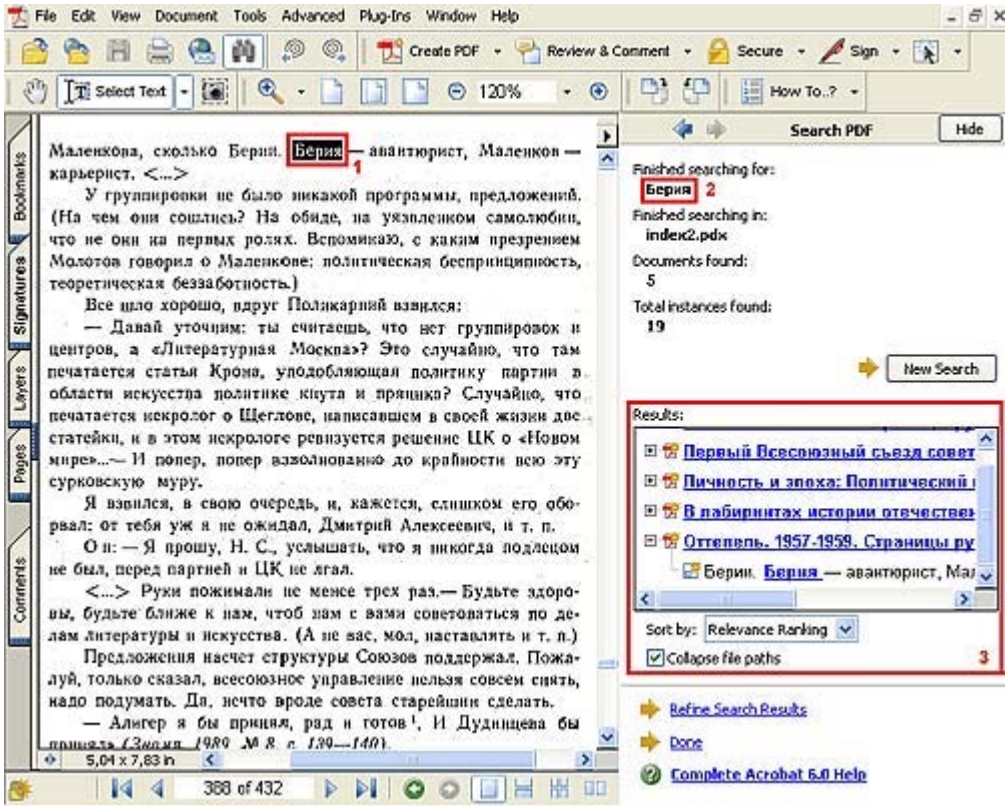
Всего было подготовлены материалы для двух учебных курсов: «История отечественной журналистики» и «Введение в литературоведение». Ежегодно эти два курса изучает около 1400 студентов пяти специальностей на трех факультетах УрГУ. Выбор именно этих учебных курсов основывался на степени «дефицитности» литературы в нашей библиотеке. Для первого курса было использовано 44 книги и 16 статей, для второго — 38 книг.

Все работы по оцифровке изданий были выполнены в течение двух летних месяцев. К работе над проектом привлекались студенты, проходившие летнюю практику в библиотеке. Грамотное построение технологического процесса позволило нам всего за два месяца оцифровать большее количество страниц, чем за предыдущие 2 года [7].

Из технологических новинок, опробованных в этом проекте, следует отметить реализацию полнотекстового поиска средствами Acrobat 6.0. Каждая электронная копия содержит «невидимый» слой распознанного текста, по которому организовывается полнотекстовый поиск. Качество распознавания обеспечивает правильное воспроизведение не менее 95% слов, что вполне достаточно для организации полнотекстового поиска. Исправление ошибок распознавания не производилось. Поиск по коллекции целиком основан на предварительном индексировании всех материалов средствами Acrobat Professional 6.0. Поисковые запросы могут быть заданы с учетом информации, находящейся в полях метаописания, что дает возможность, например, искать в нескольких изданиях одного автора или ограничить область поиска отдельной книгой. Индексный файл записывается на компакт-диск вместе с материалами коллекции. Для работы пользователя требуется всего лишь Adobe Reader 6.0, бесплатно распространяемый фирмой Adobe и включаемый нами на каждый диск с материалами коллекции. При необходимости созданные файлы могут быть проиндексированы и другими специализированными поисковыми системами.



Форма для ввода поискового запроса. Поиск по автору и заглавию возможен только при заполненных соответствующих полях PDF-документов



Результат выполнения поискового запроса:
1 — выделение найденного фрагмента подсветкой; 2 — поисковый запрос; 3 — зона навигации по результатам поиска

Создаваемые нами учебные коллекции решают конкретные задачи библиотеки:

- обеспечивают студентов «дефицитными» учебными материалами;
- снижают нагрузку на книжный фонд библиотеки и помогают его сохранять.

В то же время они открывают очень широкие возможности для самостоятельной научно-исследовательской работы студентов и аспирантов, так как в состав коллекции включены издания в полном объеме, а не фрагменты (что характерно для хрестоматий).

Защита электронных ресурсов

Определенное внимание при реализации всех наших проектов было уделено защите создаваемых полнотекстовых ресурсов. Первым уровнем защиты является ограничение выдачи цифровых копий только внутри библиотеки, для этого используется фильтрация по IP-адресам компьютеров. Второй уровень — это защита на уровне файлов цифровых копий. Все файлы в pdf-формате закрыты от изменения уникальными паролями, что должно обеспечить их неизменность. Применяются также ограничения на возможность копирования текстовой информации и распечатки файла. Степень устанавливаемых ограничений зависит от принадлежности файла к определенной цифровой коллекции. Авторы диссертаций вправе сами выбирать уровень защиты и доступа к файлу, который закрепляется в договоре. Для цифровых копий с ограниченным доступом изготавливаются демонстрационные фрагменты, которые позволяют оценить качество копии и, при необходимости, заказать ее на условиях электронной доставки документов. Электронные копии с открытым доступом располагаются таким образом, чтобы они были проиндексированы поисковыми машинами и не попали в область «невидимого веба» [8].

Сотрудничество

Побочным эффектом работ по созданию и сбору электронных ресурсов стало расширение контактов Научной библиотеки УрГУ. В августе 2003 г. с Российской Государственной библиотекой (РГБ) заключен договор о сотрудничестве в рамках проекта создания национальной электронной

библиотеки авторефератов и диссертаций. Уральский государственный университет стал первым российским университетом, заключившим такой договор с РГБ, тем самым наша библиотека вносит посильный вклад в реализацию проектов общероссийского уровня.
Определенные контакты возникли в вопросе обмена цифровыми копиями и совместной работы над тематическими коллекциями. Первый такой обмен произошел с Челябинской областной универсальной библиотекой [9]. Мы надеемся, что по мере расширения работ по самостоятельной оцифровке фондов обмен создаваемыми электронными ресурсами превратится в привычную практику для библиотек.
Несомненно, что работа над проектами по оцифровке ресурсов повысила квалификацию всех занятых в этом процессе сотрудников. Было проведено множество экспериментов, тестирований программных продуктов и технологий. Наш опыт и наши знания были востребованы: в декабре 2002 года мы провели 2 двухдневных тренинга по технологиям создания цифровых коллекций, на которых прошли обучение 30 специалистов не только из Екатеринбурга, но и из других городов Урала, Сибири и Средней России. Мы также рассказывали о своем опыте на конференциях различного уровня: от региональных до международных.
Почти 3 года работ над самостоятельным созданием и сбором цифровых полнотекстовых ресурсов и внедрением их в библиотечное обслуживание убедили нас в простой мысли: эта работа должна и может быть выгодной для библиотек. Только в этом случае она будет эффективной. И это проблема не технологическая и не финансовая, а проблема менеджмента, четкого понимания своих задач и путей их достижения.

Литература

[1] Смотри, например, доклад о деятельности Научной электронной библиотеки: Еременко Г. О. Научной электронной библиотеке — 3 года: Некоторые итоги и основные пути дальнейшего развития [Электронный ресурс] // Электронные библиотеки. — 2003. — Т. 6, вып. 1. — URL: http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2003/part1/eremenko [1 ноября 2003].
[2] Подробное описание проекта см.: Негуляев Е. А., Охезина Е. А. Цифровые коллекции в вузовской библиотеке: Концепция и технологические решения // Библиотеки и ассоциации в меняющемся мире: Новые технологии и новые формы сотрудничества: Труды 9-й Международной конференции «Крым 2002». — М., 2002. — С. 271–274. — URL: http://www.gpntb.ru/win/inte r-events/crimea2002/trud/sec4/Doc24.HTML [1 ноября 2003].
[3] Там же.
[4] Армс В. Электронные библиотеки. — М., 2001. — С. 161.
[5] Охезина Е. А. Новый интерфейс доступа к цифровой коллекции авторефератов и диссертаций [Электронный ресурс]. — 2003. — URL: http://mlist.sgu.ru/pipermail/diglib /2003-August/000048.html [1 ноября 2003].
[6] Негуляев Е. А. Yandex и полнотекстовая коллекция авторефератов [Электронный ресурс]. — 2003. — URL: http://mlist.sgu.ru/pipermail/diglib /2003-August/000050.html [1 ноября 2003].
[7] Подробнее об особенностях технологического процесса см: Негуляев Е. А. Создание электронных учебных коллекций в вузовской библиотеке // Электронные ресурсы в региональном образовании: Материалы научно-практической конференции, г. Йошкар-Ола, 8–10 октября 2003 г. — Йошкар-Ола, 2003. — С. 20–24; Негуляев Е. А., Охезина Е. А. Опыт создания цифровых учебных коллекций в вузовской библиотеке [Электронный ресурс] // Материалы VIII конференции «Электронные публикации» (El-Pub-2003), г. Новосибирск, 8–10 октября 2003 г. — URL: http://www.ict.nsc.ru/ws/show_abstrac t.dhtml?ru+76+5971 [1 ноября 2003].
[8] Негуляев Е. А. «Невидимый веб» и некоторые вопросы доступности научной информации [Электронный ресурс] // Материалы VIII конференции «Электронные публикации» (El-Pub-2003), г. Новосибирск, 8–10 октября 2003 г. — URL: http://www.ict.nsc.ru/ws/elpub2003/5972/ [1 ноября 2003].
[9] Охезина Е. А, Григорьев С. А. Первые шаги межкорпоративного сотрудничества: Создание цифровых коллекций редких книг // Информационные технологии, компьютерные системы и издательская продукция для библиотек: 5-я международная конференция и выставка LIBCOM - 2001: Тезисы докладов. — М., 2001. — С. 31–34. — URL: http://www.gpntb.ru/libcom/itog/index.cfm?n =doc/Doc14 [1 ноября 2003].